"Big Data" - Grosse Daten, viel Wissen?

Hothorn, Torsten

Abstract: Since a couple of years, the term Big Data describes technologies to extract knowledge from data. Applications of Big Data and their consequences are also increasingly discussed in the mass media. Because medicine is an empirical science, we discuss the meaning of Big Data and its potential for future medical research.

DOI: https://doi.org/10.1024/1661-8157/a001914

Other titles: "Big data" - large data, a lot of knowledge?

PRAXIS Mini-Review

Adresse:

Universität Zürich

Institut für Epidemiologie, Biostatistik und Prävention

Hirschengraben 84

CH-8001 Zürich, Schweiz

Autor: Torsten Hothorn

Haupttitel: Big Data – Grosse Daten, viel Wissen?[1]

engl. Titel: Big Data, big Knowledge?

Zusammenfassung Der Begriff *Big Data* wird seit einigen Jahren verwendet, um Technologien der empirischen Wissensgewinnung zu beschreiben und ist mittlerweile auch Diskussionsthema in den Massenmedien geworden. Da auch die Medizin eine empirische Wissenschaft ist, soll an dieser Stelle diskutiert werden, was der Begriff *Big Data* bedeutet und welcher potentielle Nutzen sich für die medizinische Forschung daraus ergibt.

Schlüsselwörter Statistik, Maschinelles Lernen, Algorithmus, Modell

[1] Dieser Aufsatz basiert auf der am 31. März 2014 gehaltenen Antrittsvorlesung von Torsten Hothorn an der Universität Zürich; ein Videomitschnitt des Vortrages ist unter `https://cast.switch.ch/vod/clips/g3to16alt` abrufbar.

1 Einleitung

Der Begriff *Big Data* geht auf einen Artikel von Chris Anderson im Wired Magazine 16.07 mit dem Titel „The End of Theory: The Data Deluge Makes the Scientific Method Obsolete“ zurück. Anderson formuliert in diesem Aufsatz die Hypothese, dass wissenschaftliche Erkenntnis zukünftig ohne Theorien auskommt und pure Korrelationen, werden sie nur aus genügend grossen Daten berechnet, die „Wahrheit“ hinreichend gut beschreiben. Tatsächlich ist die Idee nicht ganz abwegig, da uns die Wahrscheinlichkeitstheorie garantiert, dass wir aus Studien mit hinreichend grossem Stichprobenumfang auch tatsächlich die „Wahrheit“ ableiten können. Viele statistische Verfahren basieren auf mathematischen Gesetzen, welche sicherstellen, dass häufig verwendete Modelle, wie zum Beispiel die berühmte Normalverteilung, asymptotisch, also bei wachsendem Stichprobenumfang, in einem gewissen Sinne „richtig“ sind. Ob sich diese theoretische Erkenntnis und die darauf beruhenden Hoffnungen für *Big Data* auch in der Praxis, und insbesondere in der Medizin, bewahrheiten wird, soll durch einen Blick in die Geschichte näher beleuchtet werden.

2 Ein Blick zurück

Der Begriff *Big Data* steht in der Tradition einer langen Reihe von Vorgängern, welche in den letzten 60 Jahren verwendet wurden, um Methoden der empirischen Erkenntnis beschreiben. Zu nennen sind hier sind *Predictive Modelling* (Vorhersagemodellierung), *Business Intelligence* (Intelligente Geschäftsanalysen), *Machine Learning* (Maschinelles Lernen), *Artificial Neural Networks* (Künstliche Neurale Netzwerke), *Pattern Recognition* (Mustererkennung) oder *Knowledge Discovery in Data* (Datenbasierte Wissenserkennung). Allen Begriffen liegt die Idee zugrunde, dass man mit Hilfe von Computern verwertbares Wissen aus, in der Regel unstrukturierten, Datenbanken gewinnen kann. Desweiteren ist bemerkenswert, dass alle genannten Begriffe ihren Ursprung in den Computerwissenschaften haben.

Betrachtet man den *status quo* der empirischen Wissensgewinnung in der Medizin ist keiner der genannten Begriffe, vielleicht mit Ausnahme der Künstlichen Neuronalen Netze, prominent vertreten. Stattdessen besteht eine lange Tradition, klinische oder Beoachtungsstudien mithilfe von medizinstatistischen Methoden zu planen, durchzuführen, zu analysieren und zu bewerten. Es stellt sich also die Frage, wie sich die sich hinter den aus den Computerwissenschaften hervorgegangenen Begriffen stehenden Methoden zur klassischen medizinischen Statistik verhalten und welche Chancen sich daraus

für den Erkenntnisgewinn in der Medizin ergeben.

Interessanterweise beginnt Anderson seinen Aufsatz zu *Big Data* mit dem Zitat „All models are wrong, but some are useful“ des berühmten Statistikers George Box. Dahinter steht die Auffassung, dass grundsätzlich alle Modelle (und somit alle wissenschaftlichen Theorien) nur annähernd die Wahrheit (so diese denn überhaupt existiert) beschreiben und in diesem Sinne „falsch“ sind. Wenn sie aber einen Aspekt hinreichend gut und insbesondere besser als etablierte Modelle erklären, sind sie trotzdem hilfreich. In seinem kurzen Aufsatz benutzt Anderson achtmal das Wort *statistics*. Um zu verstehen, woher die Gemeinsamkeiten und Unterschiede von Statistik und *Big Data* sowie den genannten Vorgängerbegriffen kommen, ist ein Blick auf die Grundlagen der Statistik, und insbesondere der medizinischen Statistik, hilfreich.

Die Definition der Wissenschaft Statistik umfasst das Erheben, Analysieren, Interpretieren und Kommunizieren von Daten. Seinen Ursprung hat der Begriff im Wort *statisticum* (lat., den Staat beschreibend). Ursprünglich, und zu einem guten Teil bis heute, war und ist Statistik die Lehre von der Beschreibung des Staates, insbesondere seiner Bevölkerung, Wirtschaft, Verwaltung und so weiter. Mit dem Aufkommen einer mathematischen Wahrscheinlichkeitstheorie vor 250 Jahren entwickelte sich die Statistik zunehmend zu einer allgemeinen Wissenschaft der empirischen Erkenntnis, befasst sich also mit der Wissensextraktion aus Experimenten und Beobachtungen.

Ab Mitte des 19. Jahrhunderts wurden statistische Methoden sowohl in der Mathematik als auch in wichtigen Anwendungsgebieten, vor allem in der landwirtschaftlichen Forschung, der Genetik und der Medizin entwickelt. So wurde das heute unter dem Namen Fisher's exakter Test bekannte Verfahren in leicht anderer Form 1877 vom Tübinger Medizinprofessor Carl von Liebermeister entwickelt [1] und der Mediziner John Snow legte 1849 mit seiner Untersuchung der Cholera-Epidemie in London die Grundlagen der Epidemiologie.

Mit der Verbreitung von Computern in der 2. Hälfte des 20. Jahrhunderts wurden die beiden Töchter der Mathematik, die Computerwissenschaften und die Statistik, erwachsen und bildeten die heute bekannten eigenständigen Disziplinen heraus. Die Statistik ist mit der Unterteilung in Biostatistik (Biometrie) und Wirtschafts- und Sozialstatistik weiter spezialisiert. Gemeinsame methodische Grundlage ist ein hypothesengetrieber und damit modellbasierter Ansatz, in welchem zunächst eine Fragestellung mit Hilfe eines statistischen Modells formal definiert wird, dann ein entsprechendes Experiment geplant und durchgeführt wird, um schliesslich mit den gewonnenen Daten freie Parameter des Modells zu schätzen, deren Unsicherheit zu beschreiben und gegebenenfalls eine *a priori* festge-

legte Nullhypothese zu verwerfen. Der, von einem technischen Standpunkt aus betrachtet, wichtigste Punkt, die Modell- oder Parameterschätzung, das heisst, die Ableitung plausibler Modelle aus Daten, soll im Folgenden etwas näher betrachtet werden.

3 Grundlagen Statistischer Inferenz

Um statistische Modelle an Daten anzupassen geht man, stark vereinfacht natürlich, nach folgendem Prinzip vor. Zunächst formuliert man ein wahrscheinlichkeitstheoretisches Modell für das interessierende Experiment, welches unbekannte Parameter enthält. In einem zweiten Schritt definiert man eine Zielfunktion und ein daraus abgeleitetes Optimierungsproblem derart, dass man die unbekannten Parameter als eine eindeutige Lösung dieses Optimierungsproblems erhält. In einem dritten Schritt übersetzt man das Optimierungsproblem in die Praxis, indem man es mit Hilfe der experimentell gewonnenen Daten umschreibt und dann für genau dieses Experiment löst. Als Ergebnis erhält man die Parameter, welche (im Sinne des wahrscheinlichkeitstheoretischen Modells) die erhobenen Daten am besten beschreiben und interpretiert diese. Dieses theoretische Vorgehen, die sogenannte statistische Entscheidungstheorie, welche von dem Mathematiker Abraham Wald in der 1940iger Jahren entwickelt wurde, ist die gemeinsame Basis der Statistik und aller genannten Strömungen der Computerwissenschaften angefangen von den Künstlichen Neuronalen Netzwerken bis hin zu *Big Data*.

Selbst in einfachen Modellen, wie zum Beispiel einer Logistischen Regression oder einem Cox-Modell, ist es jedoch nicht mehr möglich, ein solches Optimierungsproblem mit Papier und Bleistift zu lösen. Stattdessen müssen numerische Optimierungsverfahren in einem Computer durchgeführt werden, um für erhobene Daten diejenigen Parameterwerte zu bestimmen, welche die Zielfunktion maximieren und damit die Daten am besten beschreiben. Der einzige Unterschied zwischen der Statistik und allen genannten Strömungen der Computerwissenschaften ist die genaue Wahl der Zielfunktion und die genaue Implementation des jeweiligen Optimierungsalgorithmus.

Von einer konzeptuellen Warte aus betrachtet sind also die Begriffe Statistik und *Big Data* oder zum Beispiel *Machine Learning* äquivalent. Das zeigt sich auch in der Gleichbedeutung vieler in den verschiedenen Feldern verwendeten Begriffe. Als Beispiel soll hier eine Übersetzung von im *Machine Learning* verwendeten Begriffen in die Sprache der Statistik dienen: *supervised learning* = Regression, *target variable* = Zielgrösse, *attribute* oder *feature* = erklärende Variable oder Kovariable, *hypothesis* = Modell, *instances* oder *examples* = Beobachtungen, *learning* = Parameterschätzung und

classification = Vorhersage.

In Anbetracht dieser Gemeinsamkeiten ist es also naheliegend, nach den Unterschieden zwischen Statistik und *Machine Learning* zu fragen. Die Methoden der Statistik und des *Machine Learning* unterscheiden sich in der Wahl der Zielfunktion und deshalb auch in der Wahl des Optimierungsverfahrens. Eine sogenannte *support vector machine* für den Fall einer dichotomen Zielgrösse (im *Machine Learning*: binary classification) optimiert den sogenannten *hinge loss* während ein Logistisches Regressionsmodell in der Statistik die log-Likelihood der Binomialverteilung optimiert. In der Statistik werden oft einfach interpretierbare Modelle bevorzugt, während im *Machine Learning* auf bestmögliche Modellqualität, gegebenenfalls unter Einbeziehung von komplexen nichtlinearen Funktionen, Wert gelegt wird. Etwas provokativ könnte man sagen, dass die Computerwissenschaften eine grosse Kompetenz im Lösen komplexer Optimierungsprobleme haben während für die Statistik eher wahrscheinlichkeitstheoretische Modelle, deren Eigenschaften und deren Interpretation im Vordergrund stehen.

Geht man etwas mehr ins Detail, sieht man sehr grosse Ähnlichkeiten zwischen den beiden Disziplinen. Künstliche Neurale Netzwerke haben einen starken Bezug zur nichtlinearen Logistischen Regression, *support vector machines* und *boosting* zu Generalisierten Additiven Modellen und *decision trees* zu Regressionsbäumen. Ein Verfahren, welches in beiden Disziplinen gleichviel Aufmerksamkeit erhält, sind *random forests*. Bemerkenswert an *random forests* ist, dass der Erfinder des Verfahrens, Leo Breiman, welcher Professor für Statistik in Berkeley war, seine wegweisende Arbeit (mit mehr als 5000 Zitaten seit 2001) in der Zeitschrift „Machine Learning“ veröffentlichte. Das Verfahren wird seit mehr als 10 Jahren benutzt, um komplexe nichtlineare Fragestellungen, welche sich den klassischen Modellen entziehen, zu beantworten. Als ein Beispiel aus einer Vielzahl anderer Anwendungen sei hier die Suche nach Interaktionen von genetischen *loci*, welche in einem Zusammenhang mit speziellen Erkrankungen stehen, genannt [2].

4 Big Data

Um zur Ausgangsfrage, was eigentlich *Big Data* ist und was *Big Data* in der Medizin bedeuten kann, zurückzukehren, kann man die verschiedenen Definitionen des Begriffs frei, und zugegebenermassen etwas provokativ, übersetzen: *Big Data* meint die Anwendung klassischer statistische Methoden für die Analyse grosser Mengen ungeplant und retrospektiv erhobener Beobachtungsdaten. Es wird

impliziert, dass allein die Grösse der Datensätze völlig neue Technologien (das heisst: Optimierungsverfahren) für deren Analyse notwendig macht. Insbesondere befasst man sich im Bereich *Big Data* mit Systemen, welche die Analyse von Datenmengen die nicht mehr im Arbeitsspeicher eines handelsüblichen Rechners Platz finden, ermöglichen. Dieses Problem ist jedoch alles andere als neu und seit 300 Jahren eher die Regel als die Ausnahme. Die Statistik hat für genau diese Situation die Stichprobe in ihrem Methodenbesteck, das heisst, für den Fall dass eine Vollerhebung unmöglich ist, zieht man einfach nur eine kleine handhabbare Stichprobe und schliesst von dieser auf die Grundgesamtheit. Dieses seit mindestens 200 Jahren bekannte Verfahren scheint für *Big Data* in Vergessenheit geraten zu sein.

Im Fall von *Big Data* sind jedoch Verzerrungen, Misspezifikationen und fehlende Werte die grösseren Probleme, da die Daten in aller Regel eben nicht mittels gut geplanter Stichprobenverfahren, etwa im Rahmen randomisierter kontrollierter Studien, erhoben wurden. Dies ist auch der Grund warum Korrelationen, und seinen sie auf Abermillionen von Beobachtungen basiert, eben irreführend sein können, weil schlicht die Unabhängigkeit der Beobachtungen nicht gewährleistet ist und Verzerrungen zu erwarten sind. Entgegen der weitverbreiteten Irrmeinung, dass man mit statistischen Methoden Fehler im Design eines Experimentes *post hoc* „Herausrechnen“ kann, gilt auch für *Big Data* die Regel „*garbage in, garbage out*“.

5 Möglichkeiten grosser Datenmengen

Für eine unreflektierte Euphorie besteht also auch im Zeitalter von *Big Data* kein Grund. Man wird sich weiterhin mit Modellen und Theorien als Instrument der empirischen Erkenntnis auseinandersetzen müssen, eine simple Korrelation wird nie auch nur für einfache Fragen Erklärung genug sein.

Nichtsdestotrotz bieten Systeme zur Erfassung und Verwaltung grösserer Datenmengen auch Vorteile und eröffnen neue Möglichkeiten der Erkenntnis, auch in der Medizin. So haben grosse Datenmengen das Potential, Meta-Analysen als Methode der Wissenssynthese abzulösen. Wenn aus der Publikation einer klinischen Studie nicht nur ein Effektschätzer und ein Streuungsmass abgeleitet werden kann, wie das gegenwärtig oft der Fall ist, sondern alle relevanten Patientendaten für weitergehende Analysen zur Verfügung stehen, wird es möglich sein, eine Metastudie durch Zusammenfassung aller Patientendaten *post hoc* zu kreieren und zu analysieren. Solche *open data* Strategien werden von vielen medizinischen Fachzeitschriften verfolgt, zum Beispiel dem „New England Journal of Medici-

ne“ [3], so dass in naher Zukunft immer mehr detaillierte Patienteninformationen aus wohlgeplanten klinischen Studien zur Beantwortung neuer Fragestellungen vorhanden sein werden.

Eine schon heute verfügbare Metastudie ist die PRO-ACT Datenbank [4], welche Informationen zu mehr als 8500 an Amyotropher Lateralsklerose erkrankten Patienten aus 16 Studien zusammenführt. Diese Datenbank wurde in der *Prize4Life* Initiative [5] benutzt, um neue Biomarker zu bestimmen, welche die erwartete Geschwindigkeit der Krankheitsprogression in einem frühen Stadium der Erkrankung beschreiben [6]. Ein anderes Beispiel ist die Analyse von Daten der *European Multicenter Study about Spinal Cord Injury* mittels eines Verfahrens des *Machine Learning*, sogenannten Entscheidungsbäumen [7], um eine Regel zur Patientenstratifikation zu entwickeln, welche zukünftige klinische Studien bei Patienten mit Rückenmarksverletzungen effizienter machen wird [8].

Auf einer mehr theoretischen Ebene wird es in speziellen Situationen möglich sein, nicht nur einfache Parameter, wie zum Beispiel einen Mittelwert, gut zu schätzen, sondern auch komplexere Parameter abzuleiten. Ein Beispiel hierfür sind bedingte Verteilungen, also Generalisierungen von Regressionsmodellen. Die medizinische Statistik ist ein Vorreiter in diesem Gebiet, weil Verfahren der Überlebenszeitanalyse von jeher solche bedingten Verteilungen, wie zum Beispiel die Kaplan-Meier-Kurve oder das Cox-Modell, benutzen.

6 Schlussbemerkungen

Zugespitzt kann man sagen, dass *Big Data* ein neues, aus Marketinggesichtspunkten sehr gut gewähltes, Schlagwort in einer langen Reihe von Begriffen ist, welche auf statistischen Prinzipien beruhende Computertechniken beschreiben. Diese, vorwiegend aus den Computerwissenschaften kommenden, Techniken leisten einen wertvollen Beitrag zur Verbesserung existierender statistischer Methoden durch die Erforschung neuer Optimierungsverfahren, verschweigen aber ihren starken Bezug zur Statistik. Dies hat dazu geführt, dass der Begriff Statistik mit altbackenem Erbsenzählen assoziiert wird und nicht mit einer innovativen Wissenschaft der empirischen Erkenntnis. Selbstkritisch muss man festhalten, dass das Marketing der eigenen Disziplin extrem schlecht war und leider immer noch ist. Man bedenke nur dass es möglich war, den Begriff Biometrie, welcher seit 1945 als Name einer der führenden Zeitschriften des Faches und Name der Internationalen Biometrischen Gesellschaft etabliert war, für Fingerabdruck- und Irisscanner zu kapern.

Ein weiterer Grund für die oft abschätzige Sichtweise auf statistische Methoden und Herangehens-

weisen ist die Tatsache, dass sich insbesondere Medizinstatisiker oft in der Rolle des Wissenschaftspolizisten wiederfinden, welche in Studienkomitees streng auf die Einhaltung etablierter Standards (man denke nur an Diskussionen zu Fallzahlplanungen oder an das $p < .05$ Dogma) achten. Viele Vertreter des Faches haben diese Rolle zu sehr verinnerlicht und eine Abneigung gegen neuartige und vielleicht unorthodoxe Verfahren entwickelt und somit dieses Feld anderen überlassen. Wohin dies führt sieht man am Aufkommen der Bioinformatik vor 20 Jahren, welche im wesentlichen ein Produkt der Ignoranz der aufkommenden Molekularbiologie durch die etablierte Biostatistik in den 1990iger Jahren ist. Es sieht so aus, als würde Gleiches im Moment im Bereich der personalisierten Medizin geschehen. Von einer statistischen Warte aus betrachtet erfordert die Entwicklung von auf den Patienten zugeschnittenen Therapien Modelle, welche komplexe Interaktionen zwischen Patientencharakteristika und dem Behandlungserfolg identifizieren und beschreiben. In einem statistischen Kontext ist das gut verstanden, dieses Rad wird aber derzeit durch Compterwissenschaftler neu erfunden.

Grosse Daten sind also nicht gleichbedeutend mit viel Wissen. Eine modellbasierte und theoriegeleitete Herangehensweise ist nach wie vor unabdingbar. Fortschritt wird es dort geben, wo innovative statistische Modelle mit innovativen computerwissenschaftlichen Methoden an kleine oder grosse gut geplante Experimente angepasst werden, um hoffentlich einfache Antworten auf komplexe Fragen der Medizin zu finden. Neue Schlagworte und gutes Marketing werden dabei wenig helfen, interdisziplinäre Zusammenarbeit zwischen Medizin, Biostatistik und den Computerwissenschaften dagegen viel, wie das Beispiel *random forest* zeigt.

Bibliographie

[1] Ineichen R. Der „Vierfeldertest“ von Carl Liebermeister (Bemerkungen zur Entwicklung der medizinischen Statistik im 19. Jahrhundert). Historia Mathematica. 1994;21:28–38.

[2] Cordell HJ. Detecting Gene–gene Interactions That Underlie Human Diseases. Nature Reviews Genetics. 2009;10(6):392–404.

[3] Drazen JM. Open Data. New England Journal of Medicine. 2014;370(7):662–662.

[4] Atassi N, Berry J, Shui A, Zach N, Sherman A, Sinani E, et al. The PRO-ACT Database: Design, Initial Analyses, and Predictive Features. Neurology. 2014;Online first.

[5] Küffner R, Zach N, Norel R, Hawe J, Schoenfeld D, Wang L, et al. Crowdsourced Analysis of Clinical Trial Data to Predict Amyotrophic Lateral Sclerosis Progression. Nature Biotechnology. 2014;Accepted 2014-09-23.

[6] Hothorn T, Jung HH. RandomForest4Life: A Random Forest for Predicting ALS Disease Progression. Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration. 2014;15:444–452.

[7] Hothorn T, Hornik K, Zeileis A. Unbiased Recursive Partitioning: A Conditional Inference Framework. Journal of Computational and Graphical Statistics. 2006;15(3):651–674.

[8] Tanadini L, Steeves J, Hothorn T, Abel R, Maier D, Schubert M, et al. Identifying Homogeneous Subgroups in Neurological Disorders: Unbiased Recursive Partitioning in Cervical Complete Spinal Cord Injury. Neurorehabilitation and Neural Repair. 2014;28(6):507–515.